

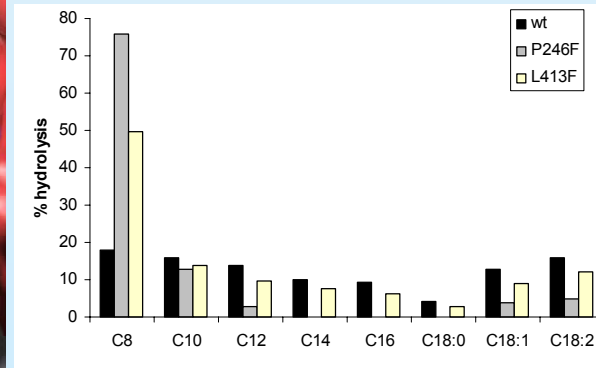
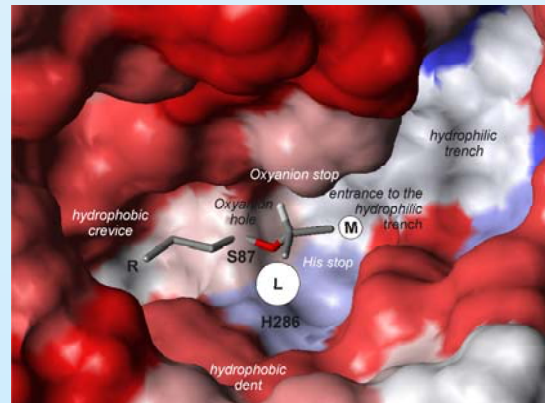
# Datenbanken in der Bioinformatik

Jürgen Pleiss

Institut für Technische Biochemie, Universität Stuttgart

**MSO**

Asp.oryzae	LKPEHSDYKIVVVGHS	LGAAIASLAAADLRT
Pen.camembertii	VVAQNPNYELVVVGH	SLGAAVATLAATDLRG
Rh.oryzae	QLTAHPTYKVIIVTGH	SLGGAQALLAGMDLYQ
1LGY-A -B -C	HHHHHHH EEEEEEE	HHHHHHHHHHHHHHHH
Rh.miehei	QFKQYPSYKVAVTGH	SLGGATALLCALDLYQ
closed (3TGL)	HHHHHHH EEEEEEE	HHHHHHHHHHHHHHHH
open (4TGL)	HHHHHHH EEEEEEE	HHHHHHHHHHHHHHHH
Hum.lanuginosa	AVREHPDYRVVFTGH	SLGGALATVAGADLRG
closed (1TIB)	HHHHHHH EEEEEEE	HHHHHHHHHHHHHHHH
Fus.heterosporum	AKTANPTFFKVVVTGH	SLGGAVATIAAAYLRK
	. . . . .	***** * * . . . *



## Datenaustausch vor 1990 ...

lokale Zentren (Bibliotheken, Rechenzentren), die lokal Daten (Publikationen, Datenbanken) zur Verfügung stellen

Kommunikation per Post:

Verschicken von Zeitschriften, Büchern, Bändern

Internet : Netzwerk von Computern mit gemeinsamem Kommunikationsprotokoll

Services: e-Mail, FTP, remote login

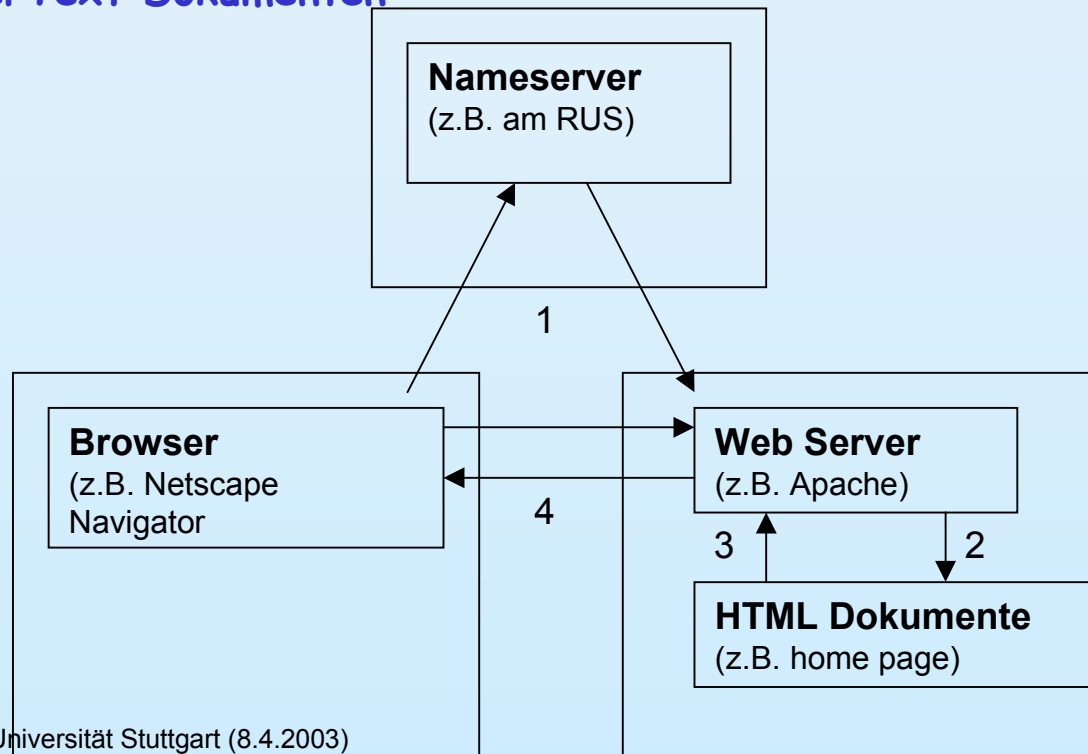
# Seit 1989: World Wide Web

1989 Tim Berners-Lee (CERN, Genf): World Wide Web

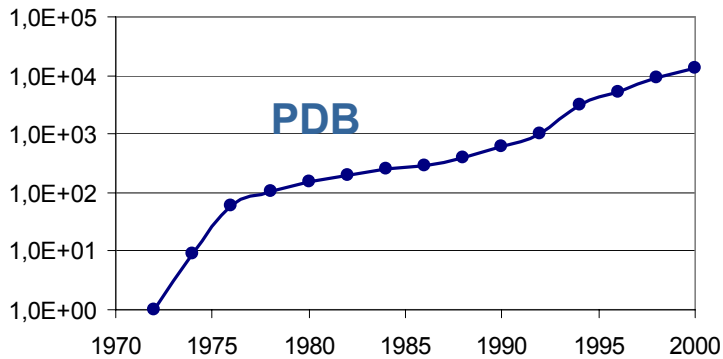
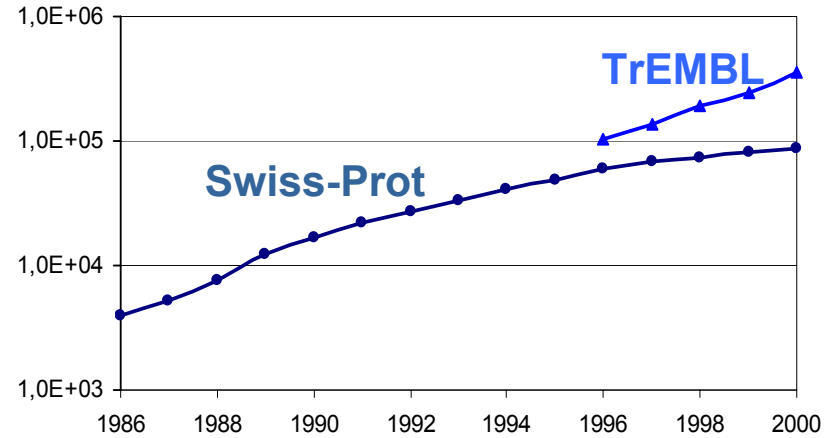
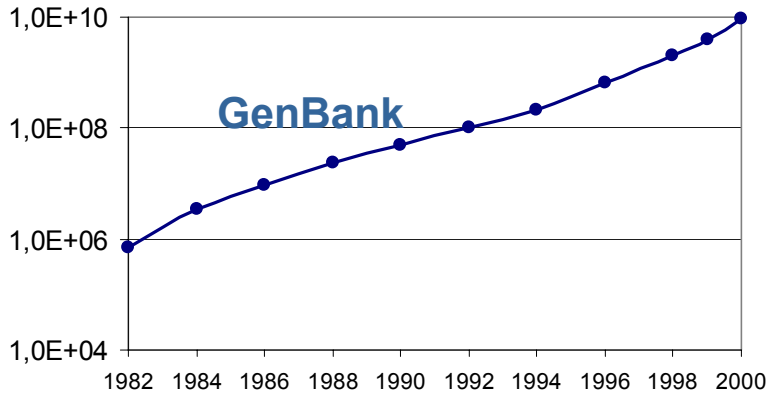
Hyper Text Markup Language (HTML)

Netzwerk von verbundenen HyperText Dokumenten

erster textbasierter Browser



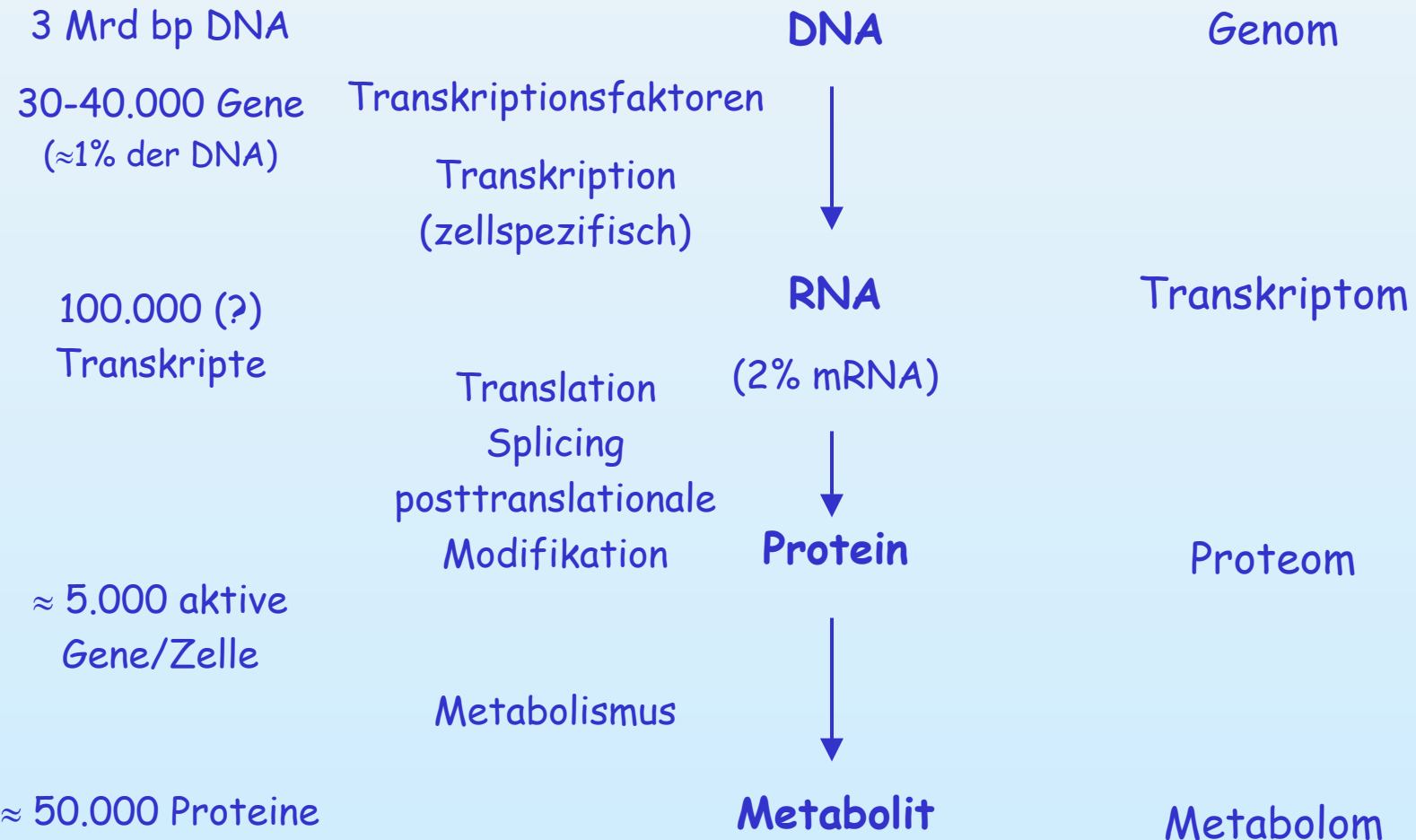
# Austausch biologischer Daten



## Oktober 2002

- 20 Mrd** Basen **+ 60%** pro Jahr
- 680.000** TrEMBL Einträge **+ 40%** pro Jahr
- 115.000** Swiss-Prot-Einträge **+ 20%** pro Jahr
- 19.000** PDB-Einträge **+ 25%** pro Jahr

# Biologische Daten: Genomik und Post-Genomik



# Was macht Bioinformatik?

## Regulation:

zeitlich

räumlich

zellspezifisch

externe Signale

Entwicklung

-omics : genomics, transcriptomics, proteomics, metabolomics

Wie funktioniert / entwickelt sich / reagiert eine Zelle / ein Organismus?

→ Funktionelle Genomik, Systembiologie

## Ein Beispiel: die *Lipase Engineering Database*

enthält

Proteinsequenz

Struktur

Informationen über einzelne Positionen

Einteilung nach Familien

Sequenzvergleich

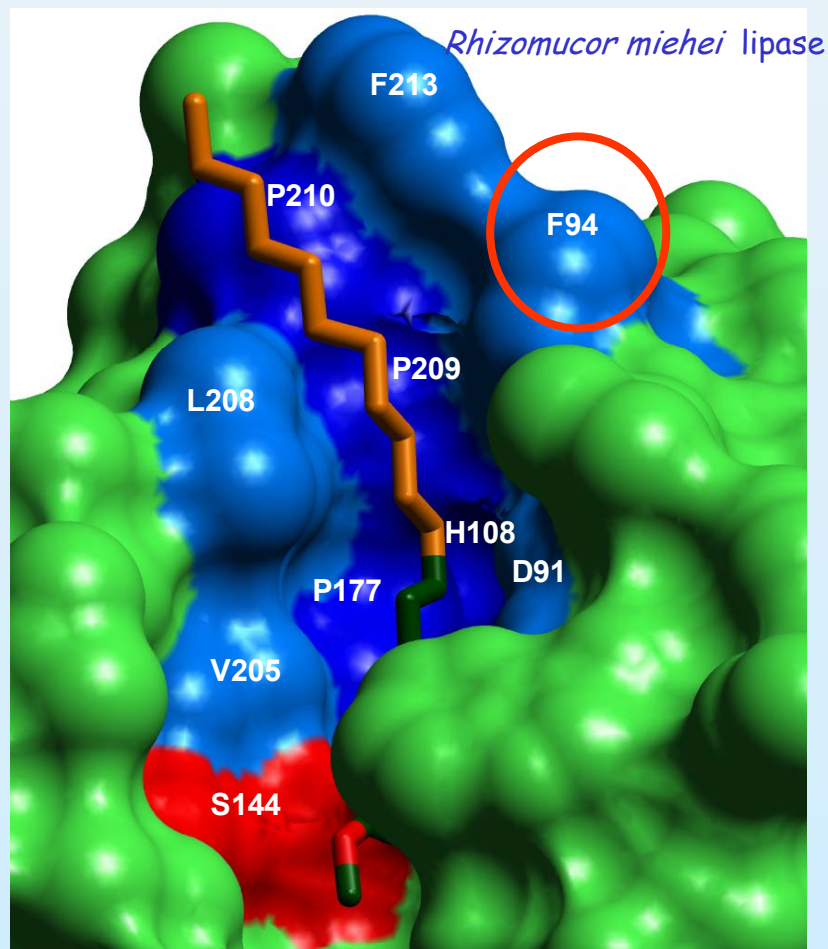
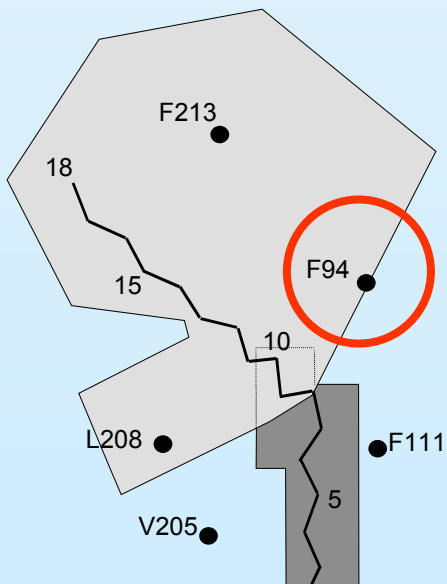
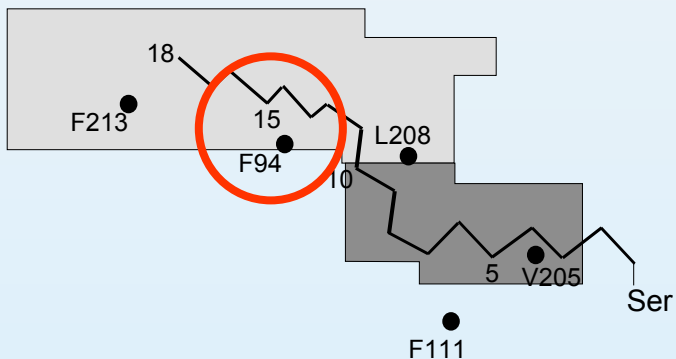
Phylogenetische Bäume

einer Enzymfamilie (>1200 Sequenzen)

Fragen: Was haben diese Enzyme gemeinsam? Was unterscheidet sie?

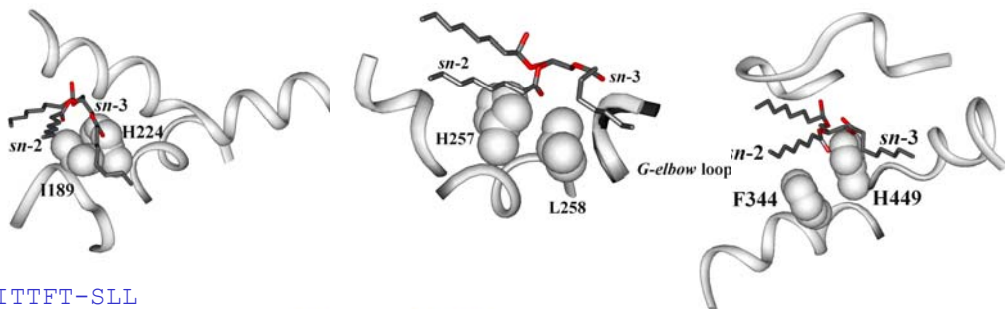
Wie kann man die Funktion in der Sequenz lesen?

# Sequenz → Struktur → Eigenschaften





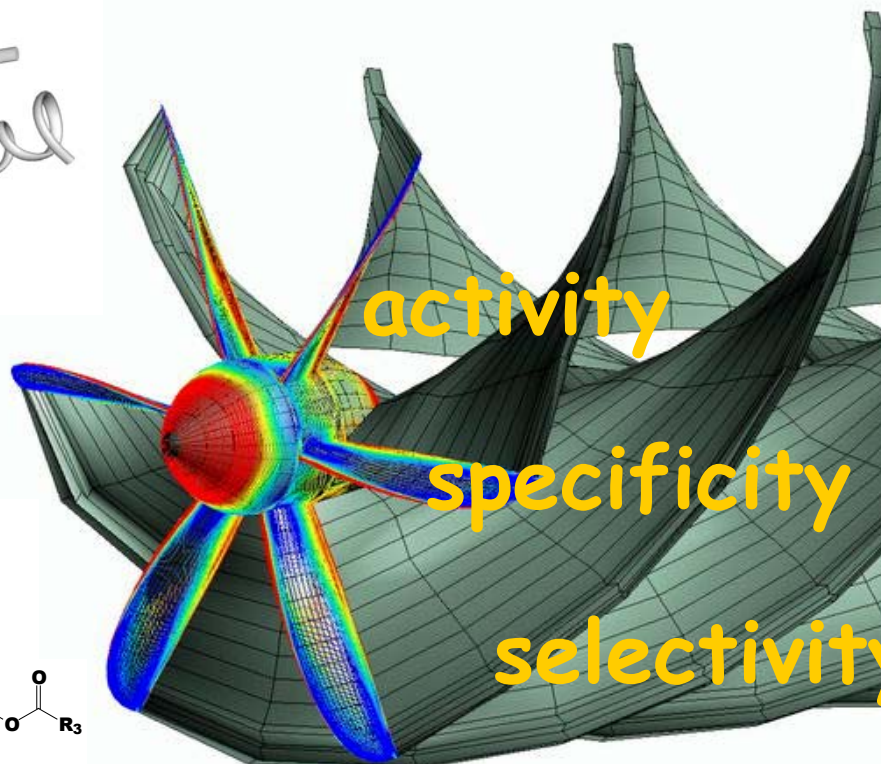
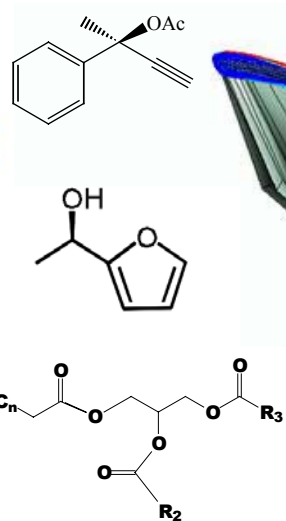
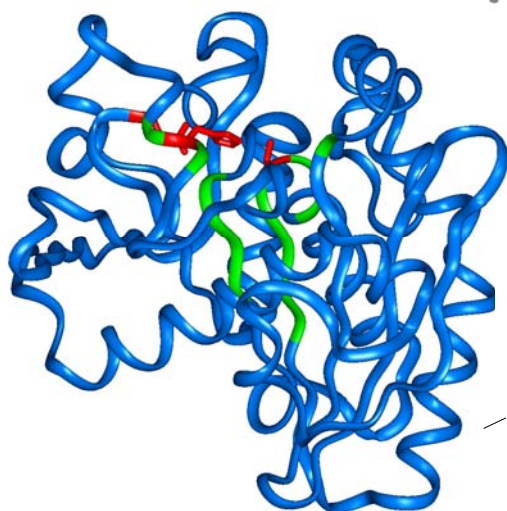
# Ziel: Verständnis der biochemischen Eigenschaften



ITTFT-SLL  
 IKTWS-TLI  
 SYDFSSTI  
 LYSFEDSGV  
 : : :

TNGYVLRSDK  
 TNAMVARGDS  
 TAGYIAVDHT  
 VTGFLALDNT  
 . . .

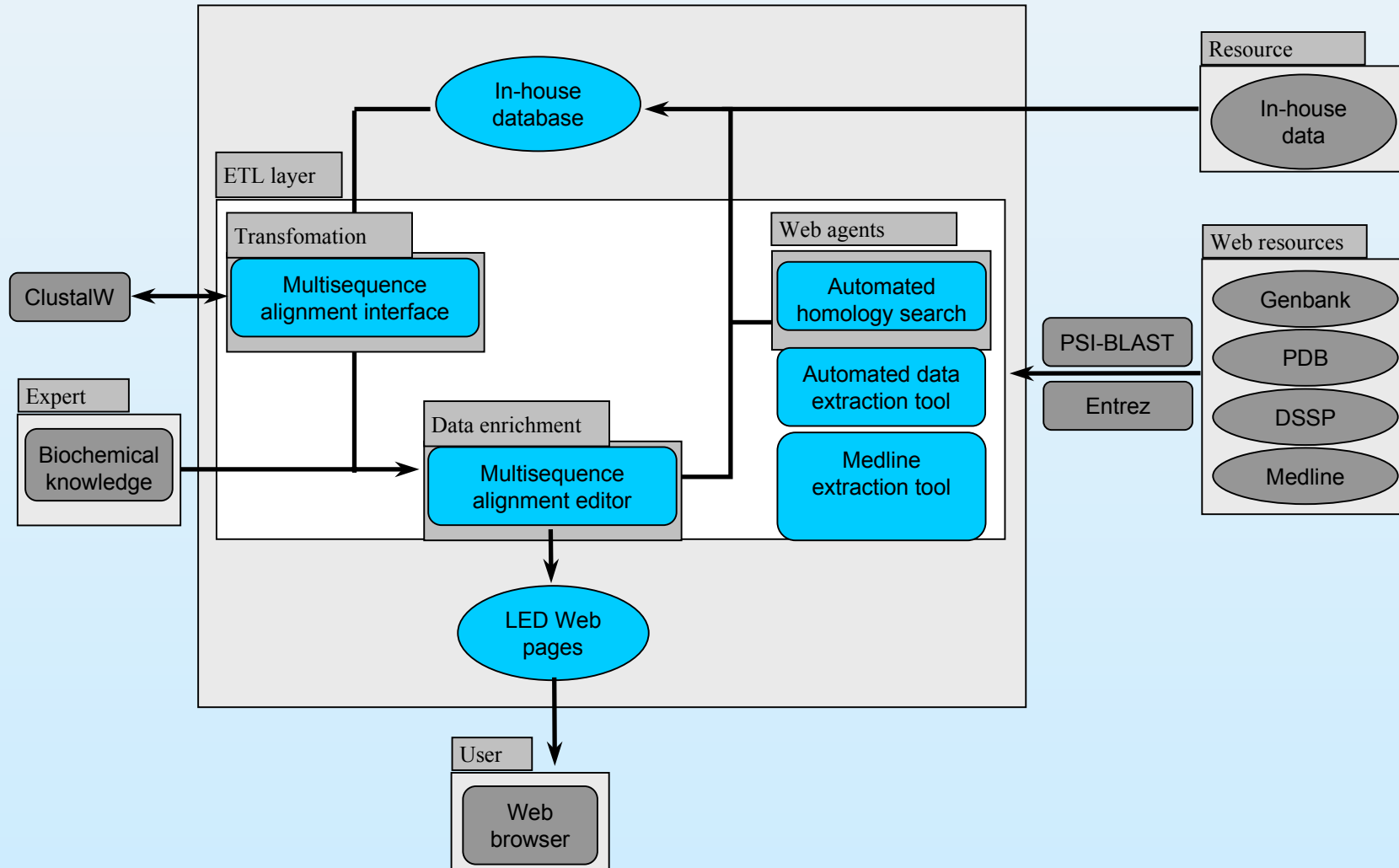
TIYLVFRGTNSF  
 TIYIVFRGSSSI  
 AVVLAFRGSYSV  
 LIVLSFRGSRSI  
 : : \*\*\* : \*



<http://www.cira.it/research/VIS/Gallery/index.htm>

GFLSSYEQVVDYFPVQEQQLTAHPTYKVIIVTGHSLGGAQALLAGMDLYQREPRAKLYAYASPRVGNAAALAKYITAQ--GNNFRFHTN-DPVPKLPILLSMGYVHVSPEYWITS  
 GFLDSYGEVQNELVATVLDQFKQYPSYKVAVTGHSLGGATALLCALDLYQREEGLSIFTVGGPRVGNPTFAYYVEST--GIPFQRTVHKRDIVPHVPPQSFGLHHPGVESWIKS  
 GFTSSWRSVADTLRQKVEDAVREHPDYRVVFTGHSLGGALATVAGADLR-----LFLYTQGGQPRVGDPAFANYVVST--GIPYRRTVNERDIVPHLPPAAFGFLHAGEEYWITD  
 GFWSSWKLVRDDIIKELKEVVAQNPNYELVVVGHSLGAAVATLAATDLR-----IDVFSYGAPRVGNRAFAEFLTVQTTGGTLYRITHTN-DIVPRLPPREFGYSHSSPEYWIKS  
 \*\* .\*: \* : : : \* \*: : .\*\*\*\*\* \* \* .. \*\* : : : .\*\*\*: : \* : : \* : \* \*\*:\* :\*: \* . \* : :

# Lipase Engineering Database : Design

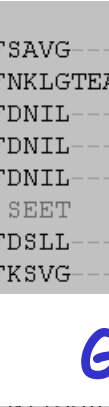


## Browse classification for GX class

Superfamilies	11
Homologous families	22
Proteins	376
Sequences	600
Structures	124
Links	<a href="#">Structure table</a>

Superfamily	Homologous family	Links
<b>Bacillus lipases</b> Proteins 6 Sequences 9 Structures 2	<a href="#">Bacillus lipases</a>	<a href="#">Alignment Tree</a>
<b>Burkholderia lipases</b> Proteins 41 Sequences 68 Structures 15 <a href="#">Alignment Tree</a>	<a href="#">Burkholderia lipases</a>	<a href="#">Alignment Tree</a>
	<a href="#">Staphylococcus lipases</a>	<a href="#">Alignment Tree</a>
<b>Candida antarctica lipase</b> Proteins 1 Sequences 8 Structures 7	<a href="#">Candida antarctica lipase B</a>	<a href="#">Alignment Tree</a>
<b>Cutinases</b> Proteins 21 Sequences 71 Structures 44 <a href="#">Alignment Tree</a>	<a href="#">Botryotinia cutinases</a>	<a href="#">Alignment Tree</a>
	<a href="#">Colletotrichum cutinases</a>	<a href="#">Alignment Tree</a>
	<a href="#">Fusarium cutinases</a>	<a href="#">Alignment Tree</a>
	<a href="#">Mycobacterium cutinases</a>	<a href="#">Alignment Tree</a>
<b>Filamentous fungi lipases</b> Proteins 36 Sequences 70 Structures 29 <a href="#">Alignment Tree</a>	<a href="#">Rhizomucor lipases</a>	<a href="#">Alignment Tree</a>
	<a href="#">Saccharomyces lipases</a>	<a href="#">Alignment Tree</a>

AAD29441	SYAKTKYPILLAHGMAGFSAVG	PLQYWNGITEDLVNGANVFVAQQASFNSSEV	RGEQLLLQAKQVLAITGAQ	128
S61927	DYAKTKYPIVLSHGLGFGNKLGTAEAFGLDYWYQIPQDLARNGANVWVTRQSTANTSEF	RGEQLLAEVQDILAITGAQ	111	
P26876	TYTQTKYPIVLAHGMGFDNILL	GVDYWFHGIPSAALRRDGAQVYVTEVSQLDTSSEV	RGEQLLQQVEEIVALSQGP	100
P26877	TYTQTKYPIVLAHGMGFDNILL	GVDYWFHGIPSAALRRDGAQVYVTEVSQLDTSSEV	RGEQLLQQVEEIVALSQGP	100
1EX9	TYTQTKYPIVLAHGMGFDNILL	GVDYWFHGIPSAALRRDGAQVYVTEVSQLDTSSEV	RGEQLLQQVEEIVALSQGP	74
1EX9A	TT SS EEEE TT SEET	TEESSTTHHHHHHHTT EEEE SSS HHH	HHHHHHHHHHHHHHHH S	
AAM14701	GYTATKYPIVLTHGMGFDSSL	GIDYWYGIPSAALRRDGAQVYITEVSQLNTSEL	RGEELLAQVEEIVAISGKP	101
G83044	DYTRTRYPIVLSHGLGFGKSVG	PVDYWHAIVPALEKDGAKVFATSSQSPVNSNEV	RGEQLLAQVEEVLALTGAE	98
AAD22078	GYTETRYPLVLVHGLGFGI	IPHALSKDGATVFTAQVAAANRSEV	RGEQLLAQVETILALTGKE	100
AAG47649	GYTETRYPLVLVHGLGFGI	IPHALSKDGATVFTAQVAAANRSEV	RGEQLLAQVETILALTGKE	100
P15493	GYTQTRYPIVLSHGLGFGI	IPQSLTRDGAQVYVAQVSATNSER	RGEQLLAQVESLLAVTGAK	102
AAB53647	NYTKTKYPIVLVHGLGFGI	IPWNLERDGARVHVASVAAFNDSEQ	RGAELARQIVPWAAGGGG	95
AAC15585	TTATRYPLVLVHGMGFGIRLLL	YPYWYGIKALRRGGATVIAVQVSPVNSTEV	RGEQLLARIDEILRETGAA	75
CAC07191	VNTRYPIILLVHGLGFGDRIGS	HHYFHGIKQALNECGASVFVPIISAANDNEA	RGDQLLKQIHNLRQVGAQ	75
AAB01071	MSTTYPIVLSHGLGFGDDIVG	YPYFYGIRDALDKGKVFATSLAFNSNEV	RGEQLWEFVQVKLKETKAK	74
3LIP	YAATRYPIILLVHGLGFGTDKYAG	VLEYWYGIQEDLQQRGATVYVANLSGFQSDDGPNRGEQLLAYVKTVLAATGAT	79	
3LIP	TT SS EEEE TT SEETT	TEESSTTHHHHHHHTT EEE SS SSSTTSHHHHHHHHHHHHHHHHT S		
AAC05510	YATTRYPIILLVHGLGFGTDKYAG	VLEYWYGIQEDLQQRGATVYVANLSGFQSDDGPKGRDEQLLAYVKTVLAATGAT	119	
P25275	YAATRYPIILLVHGLGFGTDKYAG	VVEYWYGIQEDLQQRGATVYVANLSGFQSDDGANGRGEQLLAYVKTVLAATGAT	123	
1TAH	YAATRYPVILLVHGLGFGTDKDFAN	VVDYWYGIQSDLQSHGAKVYVANLSGFQSDDGPNRGEQLLAYVQVLAATGAT	79	
1TAHD	TT SS EEEE S S TTS	TTSSTTHHHHHHHTT EEEE TT SSSTTSHHHHHHHHHHHHHHHHT S		
1CVL	YAATRYPVILLVHGLGFGTDKDFAN	VVDYWYGIQSDLQSHGAKVYVANLSGFQSDDGPNRGEQLLAYVQVLAATGAT	79	
1CVL	TT SS EEEE S S TTS	TTSSTTHHHHHHHTT EEE S S SSSTTSHHHHHHHHHHHHHHHHT S		



GX

*Candida rugosa lipases*

P32948	QMNPLGNWSSSLPKAAINSLMQS	KLFQAVLP	NGEDCLTINVRPSPGTPKANLPVMVWI	GGGF	EVGGSSLFPP	150
P32946	QMNPMGSFEDTLPKARHLVLQS	KIFQVVLN	NDEDCLTINVIRPPGTRASAGLPVMLWI	GGGF	ELGGSSLFPG	149
P20261	QQNPEGTYEENLPKAALDLVMQS	KVFEAVSP	SSSEDCLTINVRPPGPKAGANLPVMLWI	GGGF	EVGGTSTFPP	150
1LPP	QQNPEGTYEENLPKAALDLVMQS	KVFEAVSP	SSSEDCLTINVRPPGPKAGANLPVMLWI	GGGF	EVGGTSTFPP	135
1LPP	TT SS HHHHHHHHHTS	HHHHHS B S EEEEE	TT TT EEEEE	STTS	GGGS	
P32947	QQNPEGTFEENLGKTALDLVMQS	KVFQAVLP	QSEDCLTINVRPPGPKAGANLPVMLWI	GGGF	EIGSPTIFPP	150
1CLE	QQNPEGTFEENLGKTALDLVMQS	KVFQAVLP	QSEDCLTINVRPPGPKAGANLPVMLWI	GGGF	EIGSPTIFPP	135
1CLEB	TT SS HHHHHHHHHHS	HHHHHS B S EEE	TT EEE	STTS	GGGS	
P32949	QQNPEGTYEENLPKVALDLVMQS	KVFQAVLP	NSEDCLTINVRPPGPKAGANLPVMLWI	GGGF	EIGSPTIFPP	150
1THG	QLDPGNSLTLDDKALGLAKV IPEEFRGPLYDMAKGTVSMNED	CYLYLN	VWI	GGGF	VYGS SAAYPG	143
1THG	HHHHHHHHHHHH HHHHS HHHHHHHHHT S B S	EEEEETT TT EEEEE	TT SGGG	S		
BAA19072	QLDPGNSLTLDDKALGLAKV IPEEFRGPLYDMAKGTVSMNED	CYLYLN	VFRPAGTKPDAKLPVMVWI	GGGF	VYGS SAAYPG	162
P22394	QLDPGNSLTLDDKALGLAKV IPEEFRGPLYDMAKGTVSMNED	CYLYLN	VFRPAGTKPDAKLPVMVWI	GGGF	VYGS SAAYPG	162
P17573	QLDPGNAISLLDKVVLGKIIIPDNLRGPLYDMAQGSVSMNED	CYLYLN	VFRPAGTKPDAKLPVMVWI	GGGF	VYGS SASYPG	162
S59957	QLNPGNALTILDNALSIS ISENIRGPLYDMAKGSVSMSEDC	CYLYN	VCRPAGTKPGDKLPVMVWI	GGGF	VYGS SRSYPG	162
S	* * . . . . .	****	** ** **	. . . . .	****	** ** * . *



GGGX

**Lipoprotein lipases**

Hepatic lipase Pancreatic

Moraxella lipase 3

Mycoplasma lipase

Non-heme peroxidase

**Haemophilus influenzae**

Haemophilus

**Gastric lipases**

# Vielfalt von Sequenz ...

**Cutinases**

Mycobacterium tuberculosis  
Colletotrichum gloeosporioides  
Fusarium solani

Staphylococcus lipase  
Cepacia lipase

**Candida antarctica**

Candida antarctica lipase B

**Bacillus subtilis**

Bacillus lipase

**Moraxella1**

Moraxella lipase 1

**Pseudomonas fluorescens**

Pseudomonas fluorescens lipase

**Acinetobacter calcoaceticus**

Acinetobacter esterase

**Moraxella2**

Moraxella lipase 2

**Yarrowia lipolytica**

Yarrowia lipolytica lipase

**Carboxylesterases**

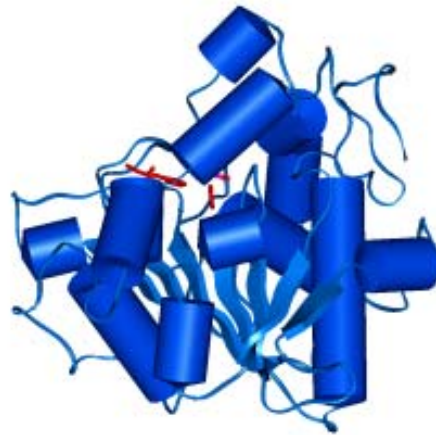
**Candida rugosa**

Bacillus subtilis  
Candida rugosa lipase  
Arthrobacter oxidans  
Pseudoobscura esterase  
Dictyostelium esterase  
Mycus persicae  
Mammalian carboxylesterase  
Mammalian bile salt activated lipase  
Torpedo californica  
Ca.elegans cytoplasmic esterase  
Heliothis virescens  
Culex pipiens

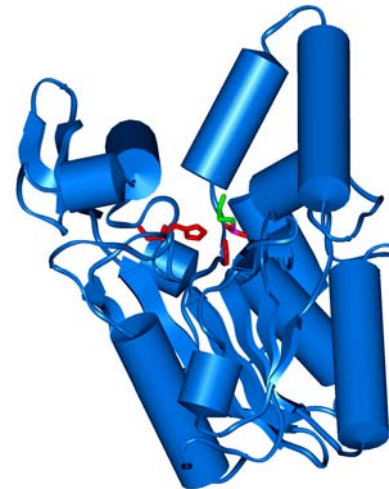
## ... und Struktur



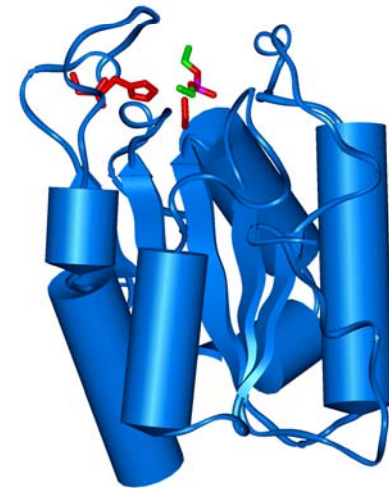
*Rhizomucor miehei* lipase



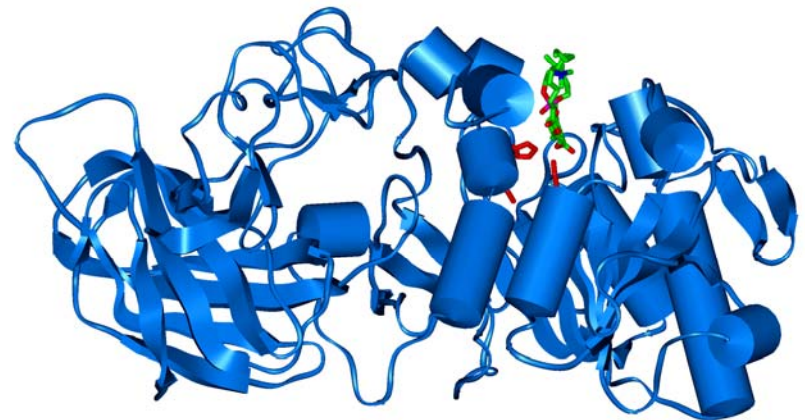
*Candida antarctica* lipase B



*Burkholderia cepacia* lipase



*Fusarium solani* cutinase



Human pancreas lipase

# Sequenzdatenbank: GenBank

Entrez-Nucleotide - Netscape

File Edit View Go Communicator Help

NCBI

CGCTCAGGATAGGACTTCGGTCGCTAGAGGATCGGATCCCCGGCGGATTATATAGCTCGATCGATC1  
TTCTCTATATCCGCGGATAGGGCTATATACACACACAGCCCGCGGATAGCATGACTGATCTA  
CCCCAATGACCTGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGAT  
CACAGACTGACGCTGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGAT

Entrez Nucleotide

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM

Search Nucleotide for lipase AND rhizopus

Clear

Limits Preview/Index History Clipboard Details

About Entrez

Search for Genes  
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

The Entrez Nucleotides database is a collection of sequences from several sources, including GenBank, RefSeq, and PDB. The number of bases grows at an exponential rate. Today's total is:

18025031554

Document: Done

Suchbegriffe

Zugang über Entrez : <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide>

# Sequenzdatenbank: GenBank

The screenshot shows the NCBI Sequence Viewer interface. The 'FEATURES' section includes:

- source: 1.1600, /organism="Rhizopus niveus", /db\_xref="taxon:4844"
- CDS: 236..1414, /codon\_start=1, /product="lipase", /protein\_id="BAAD2493.1", /db\_xref="GI:218043", /translation="MVSFISISQGVSLCLLVSSMMLGSSAVPVSGKSGSSNTAVSASD NAALPPLISSRCAPPNKGSKSDLQAEPYNMQKNTWYESHGGLNLTIGKRDDNLVGG MTLDLPSDAPPISLSSSTNSASDGGKVVAAATTAQIQEFTKYAGIAATAYCRSVVPGNK WDCVQCQKWVPDGKIITTFSSLSDINGVLRSDRQKTIYLVFRGINSFRSAITDIV NFDYKPKVKGAKVHAGFLSSYEQVVDYFVQVEQLTAHPTYKVIVTGHSLGGAQALL AGMDLYQREPRPLSKNLSIFTVGGPRVGNPTFAYYVESTGIPFQRTVHKRDIVPHVPP QSFGLHPGVESWIKSGTSNVQICTSEIETKDCSNSIVPFTSILDHLSYFDINEGSCL "

The 'BASE COUNT' section shows: 400 a, 395 c, 278 g, 527 t.

The 'ORIGIN' section displays the DNA sequence in 10-line blocks:

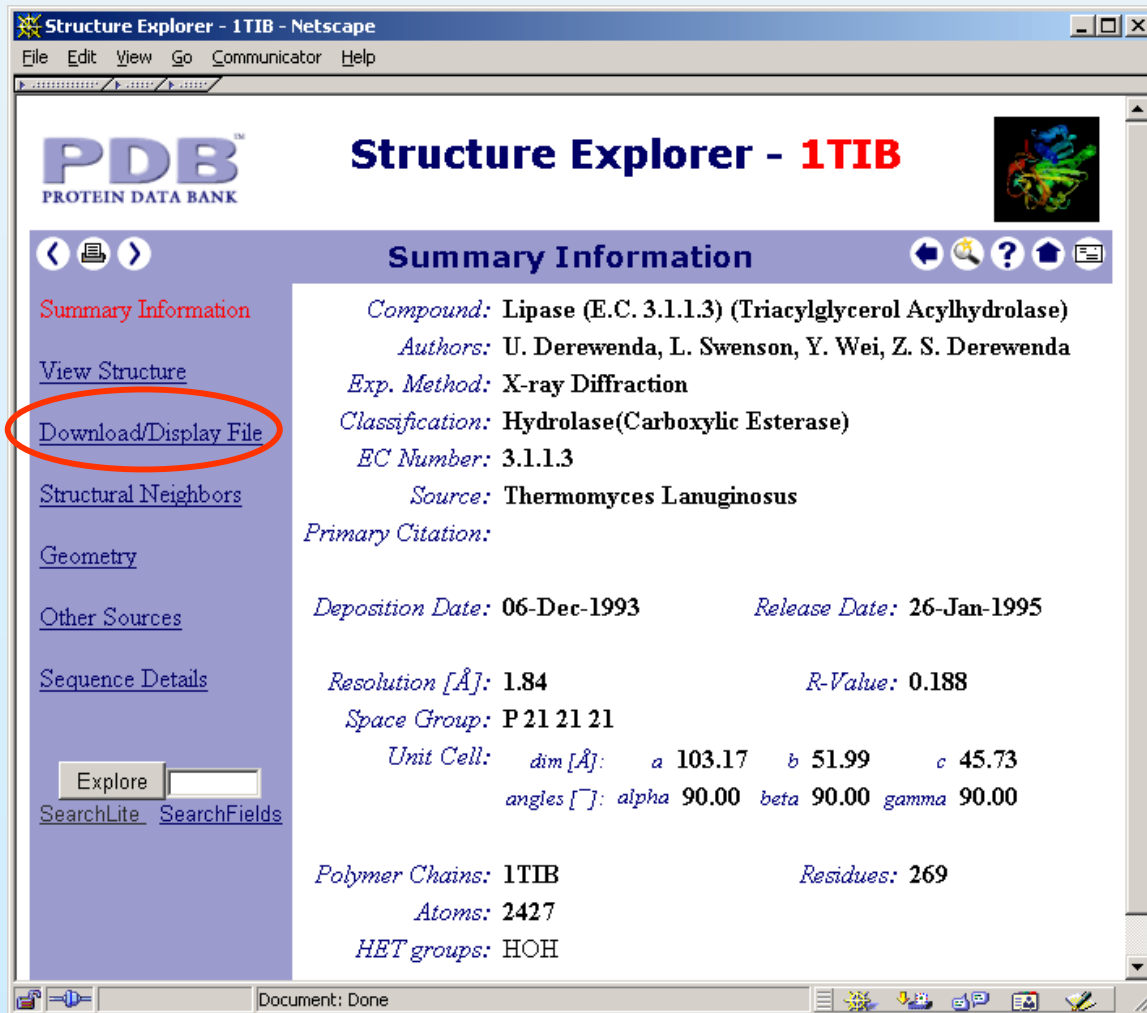
```
1 aatcagtcgt aacaataatt gattacttgg tactactatt aaatgtacct aatttcatga
61 ggggttacaa tgtgcgtgga taaattgcca ttggtctcgc tattttttga aaaaaaaaaa
121 catataaata gaggccagtt tatggtatgt tcaagtctct atcttcatca agtcaaagtg
181 atacagactc ttcttttctt ttcttcttac cccttccagt tctttactat caatcatggt
241 ttcattcatt tccatttctc aaggtgtag tctttgtctt cttgtctctt ccgatgatgc
301 cggttcatct gctgttctctg ttcttggtaa atctggatct tccaacaccg ccgtctctgc
361 atctgacaat gctgccctcc ctctctctcat ctccagccgt tgtgctctcc ctcttaacaa
421 gggaaagtaa agcgatctcc aagctgaacc ttacaacatg caaaagaata cagaatggta
481 tgagtccccat ggtggcaacc tgacatccat cggaaagcgt gatgacaact tggttggtgg
541 catgactttg gacttaccac gcgatgctcc tctatcagc ctctctagct ctaccaacag
601 cgcctctgat ggtgtaaggt ttgttgcctc tactactgct cagatccaag agttcaccac
661 gtatgctggt atcgcctgcca ctgcctactg tcgttctggt gtcctcggtc acaagtggga
721 ttgtgtccaa tgtcaaaagt gggttcctga tggcaagatc atcactacct ttacctcctt
781 gctttccgat acaaatggtt acgtcttgag aagtgataaa caaaagacca tttatcttgg
841 ttteccgtggt accaactctc tcagaagtgc catcactgat atcgtctcca acttttctga
901 ctacaagcct gtcaagggcg ccaaagtcca tgetggttcc ctttctctct atgagcaagt
961 tgtcaatgac tatttccctg tcgtccaaga acaattgacc gcccacccta cttataaggt
1021 catcgttacc ggtcactcac tgggtggtgc acaagctttg cttgcccgtc tggatctcta
1081 ccaacgtgaa ccaagattgt ctcccaagaa tttgagcacc ttcactgtcg gtggctctcg
1141 tgttggtaac cccacctttg ctactatgt tgaatccacc ggtatccctt tccaacgtac
1201 cgttcacaag agagatatcg tctctcaagt tctctctcaa tctctcggat tctctcatcc
1261 cgggtgtgaa tcttggatca agtctggtac ttccaacgtt caaatctgta cttctgaaat
1321 tgaaccaaac gattgcagta actctatcgt tcttttccac tctatccttg accacttgag
```

Coding sequence

DNA-Sequenz



# Strukturdatenbank: Protein Data Bank



**PDB**  
PROTEIN DATA BANK

## Structure Explorer - 1TIB

**Summary Information**

**Compound:** Lipase (E.C. 3.1.1.3) (Triacylglycerol Acylhydrolase)  
**Authors:** U. Derewenda, L. Swenson, Y. Wei, Z. S. Derewenda  
**Exp. Method:** X-ray Diffraction  
**Classification:** Hydrolase(Carboxylic Esterase)  
**EC Number:** 3.1.1.3  
**Source:** Thermomyces Lanuginosus

**Primary Citation:**

**Deposition Date:** 06-Dec-1993      **Release Date:** 26-Jan-1995

**Resolution [ $\text{\AA}$ ]:** 1.84      **R-Value:** 0.188  
**Space Group:** P 21 21 21

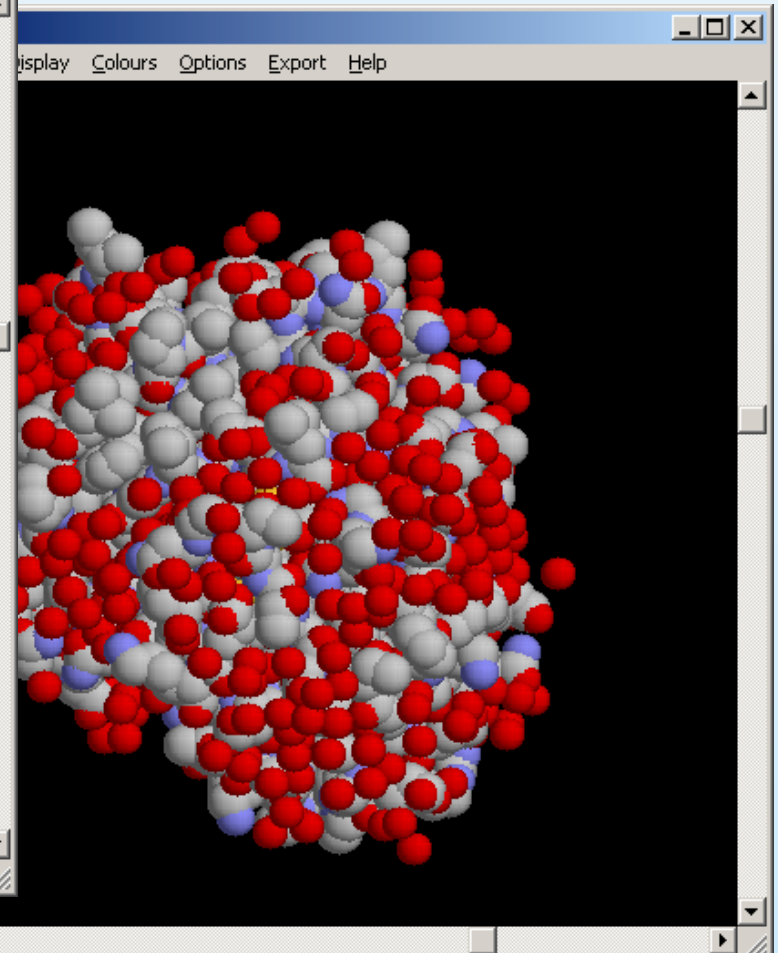
**Unit Cell:**    *dim [ $\text{\AA}$ ]:*    *a* 103.17    *b* 51.99    *c* 45.73  
*angles [ $^\circ$ ]:*    *alpha* 90.00    *beta* 90.00    *gamma* 90.00

**Polymer Chains:** 1TIB      **Residues:** 269  
**Atoms:** 2427  
**HET groups:** HOH

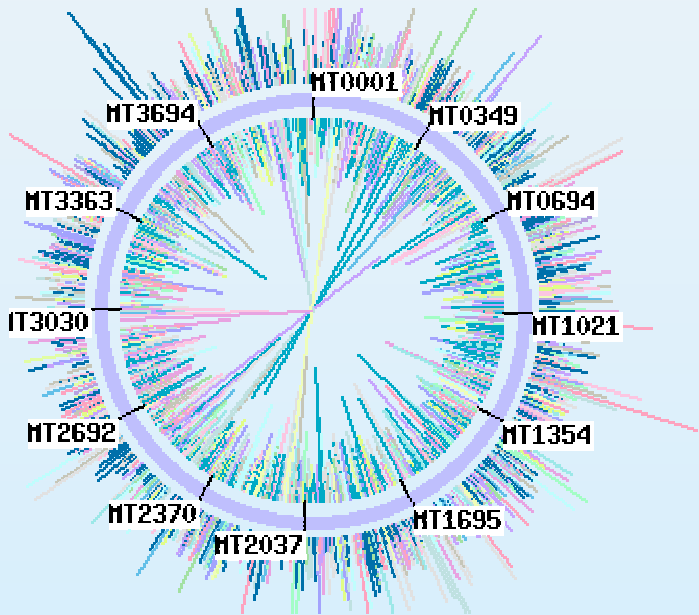
Explore

[SearchLite](#) [SearchFields](#)

# Visualisierung von Strukturen



# Genomics Datenbanken



Summary sequence view - Mozilla

## Myco**ba**cterium tuberculo**si**s CDC1551, complete genome - 1740750..1790749

49 protein coding genes  Find Open Reading Frames

Click on the rectangle to get BLAST neighbors for the gene of interest or click on the overview below to see a distant region

1740750 1742882 1745015 1747148 1749280

1750750 1752882 1755015 1757148 1759280

1760750 1762882 1765015 1767148 1769280

1770750 1772882 1775015 1777148 1779280

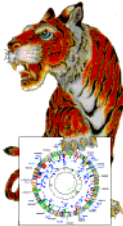
1780750 1782882 1785015 1787148 1789280

- Translation, ribosomal structure and biogenesis
- Transcription
- DNA replication, recombination and repair
- Cell division and chromosome partitioning
- Posttranslational modification, protein turnover
- Cell envelope biogenesis, outer membrane
- Cell motility and secretion
- Inorganic ion transport and metabolism
- Signal transduction mechanisms
- Energy production and conversion
- Carbohydrate transport and metabolism

# Genomics Datenbanken

TIGR Microbial Database - Netscape






File Edit View Go Communicator Help



## TIGR Microbial Database

a listing of published microbial genomes and

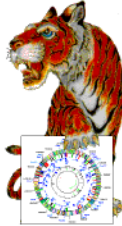
Published complete microbial genomes (listed a

Link	Genome	Strain	Domain	Size (Mb)	Institut
	<i>Aeropyrum pernix</i>	K1	<a href="#">A</a>	1.67	<a href="#">Biotechn Cent</a>
	<i>Aquifex aeolicus</i>	VF5	<a href="#">B</a>	1.50	<a href="#">Diver</a>
	<a href="#">Archaeoglobus fulgidus</a>	DSM4304	<a href="#">A</a>	2.18	<a href="#">TIGR</a>
	<i>Bacillus subtilis</i>	168	<a href="#">B</a>	4.20	Internat Consor
	<a href="#">Borrelia burgdorferi</a>	B31	<a href="#">B</a>	1.44	<a href="#">TIGR</a>

Document: Done

TIGR Microbial Database - Netscape

File Edit View Go Communicator Help



## TIGR Microbial Database:

a listing of microbial genomes and chromosomes in progress

Microbial genomes and chromosomes in progress (Searches available for some TIGR genomes)


Genome	Strain	Domain	Size (Mb)	Institution	Funding	Anticipated Completion
<i>Actinobacillus actinomycetemcomitans</i>	HK1651	<a href="#">B</a>	2.2	<a href="#">University of Oklahoma</a>	<a href="#">NIDR</a>	
<i>Agrobacterium tumefaciens</i>	C58	<a href="#">B</a>	5.3	University of Washington / Dupont	Dupont	2001
<i>Aspergillus nidulans</i>		<a href="#">E</a>	29	Cereon Genomics		
<i>Bacillus anthracis</i> <a href="#">BLAST Search</a>	Ames	<a href="#">B</a>	4.5	<a href="#">TIGR</a>	<a href="#">ONR / DOE / NIAID</a>	
<i>Bacillus halodurans</i>	C-125	<a href="#">B</a>	4.2	<a href="#">Japan Marine Science and Technology Center</a>		Complete
<i>Bacillus stearothermophilus</i>	10	<a href="#">B</a>		<a href="#">Univ. of Oklahoma</a>	<a href="#">NSF</a>	
<i>Bartonella henselae</i>	Houston 1	<a href="#">B</a>	2.00	<a href="#">University of Uppsala</a>	<a href="#">SSF</a>	2000
<i>Bordetella bronchiseptica</i>	RB50	<a href="#">B</a>	4.9	<a href="#">Sanger Centre</a>	<a href="#">Beowulf Genomics</a>	
<i>Bordetella parapertussis</i>		<a href="#">B</a>	3.9	<a href="#">Sanger Centre</a>	<a href="#">Beowulf Genomics</a>	


Document: Done

# Human Genome Project

The Sanger Institute : Human Genome Project - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

 [Sanger Home](#) | [Acedb](#) | [YourGenome](#) | [Ensembl](#) | [Trace Server](#) | [Library](#)

 The Wellcome Trust  
Sanger Institute

[Info](#) | [Databases](#) | [Blast](#) | [Genomics](#) | [Infrastructure](#) | [HGP](#) | [CGP](#) | [Projects](#) | [Software](#) | [Teams](#) | [Search](#) [Data Release Policy](#) | [Conditions of Use](#)

## Human Genome Project

**HGP Home**

Chr ?

[Annotation](#)

[Genes](#)

[Polymorphism](#)

[RH/EST Maps](#)

[Cytogenetics](#)

[CpG islands](#)

**Ensembl Genome Browser**

Chr ?

**HGP Home**

**PUBLICATION 2001**

**DRAFT 2000**

[Overview](#)

[Targets](#)

### Human Genome Project at the Sanger Institute

The human genome research programme at [The Sanger Institute](#) encompasses mapping, sequencing, and structural and functional interpretation. A general [overview](#) of the human genome project is available.

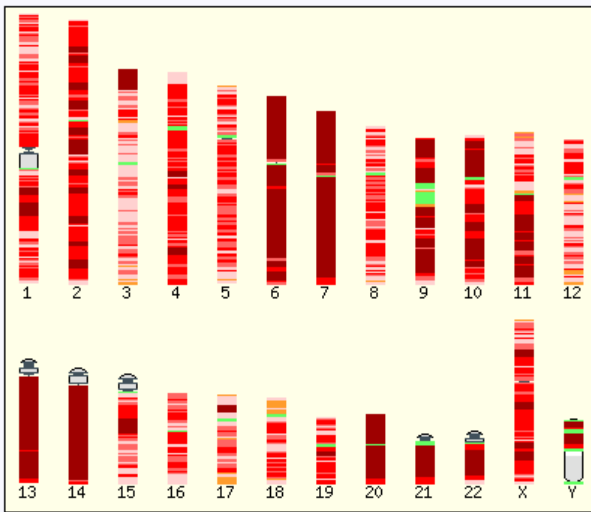
The Sanger Institute is engaged upon [collaborative projects](#) to sequence all or part of chromosomes [1](#), [6](#), [9](#), [10](#), [11](#), [13](#), [20](#), [22](#) and [X](#).

Access chromosome sequencing project information, [sequence](#) and [map](#) data by selecting a chromosome number from the sidebar or click the appropriate chromosome image.

Percentage	
Draft	11.1%
Finished	81.3%
Total	92.4%

Sequencing status

### Assembly Completion Status



Key

# Polymorphismen

SNP - Microsoft Internet Explorer

File Bearbeiten Ansicht Favoriten Extras ?

NCBI

ENTREZ **SNP**  
Single Nucleotide Polymorphism

PubMed Nucleotide Protein Genome Structure Popset Taxonomy

Search SNP for Go Clear

Limits Preview/Index History Clipboard Detail

- Enter one or more search terms.
- Available search fields are listed below
- Use [Limits](#) to restrict your search by search field, chromosome, and other criteria.

## SNP

dbSNP is now incorporated into NCBI's Entrez system and can be queried using the same approach as the other Entrez databases such as PubMed and GenBank. The original database with additional information and search options are available [here](#).

**Update:**  
August 14, 2002 Add contig position tag [CTPOS]

Below are search examples and available search fields.

Search using wild-card(\*), ranging(:), AND, OR, and NOT operators:

Microsoft Internet Explorer

sicht Favoriten Extras ?

## Single Nucleotide Polymorphism

Select the BLAST program to use and enter your sequence in the text area below.

Program **blastn**

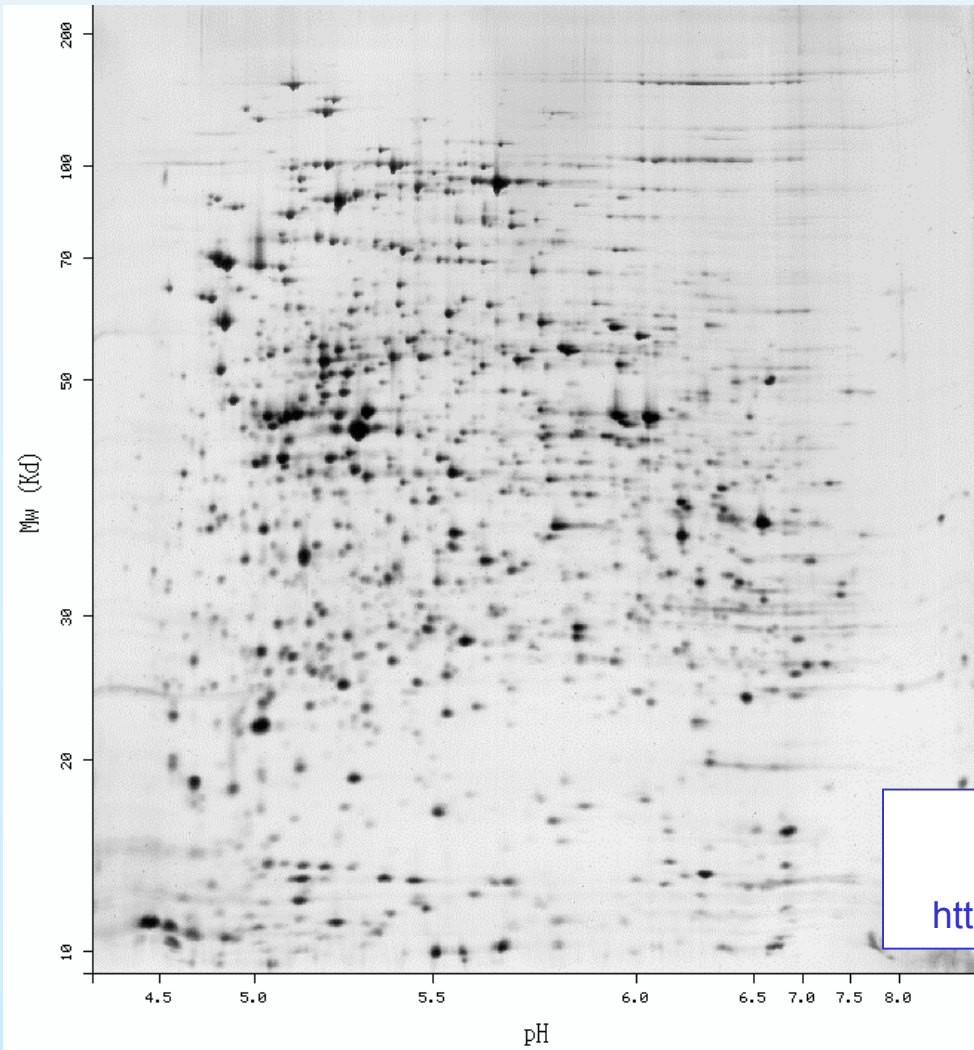
### Query Sequence

Enter an accession, gi, or a sequence in [FASTA](#) format

Anfrage senden Clear Input

Internet

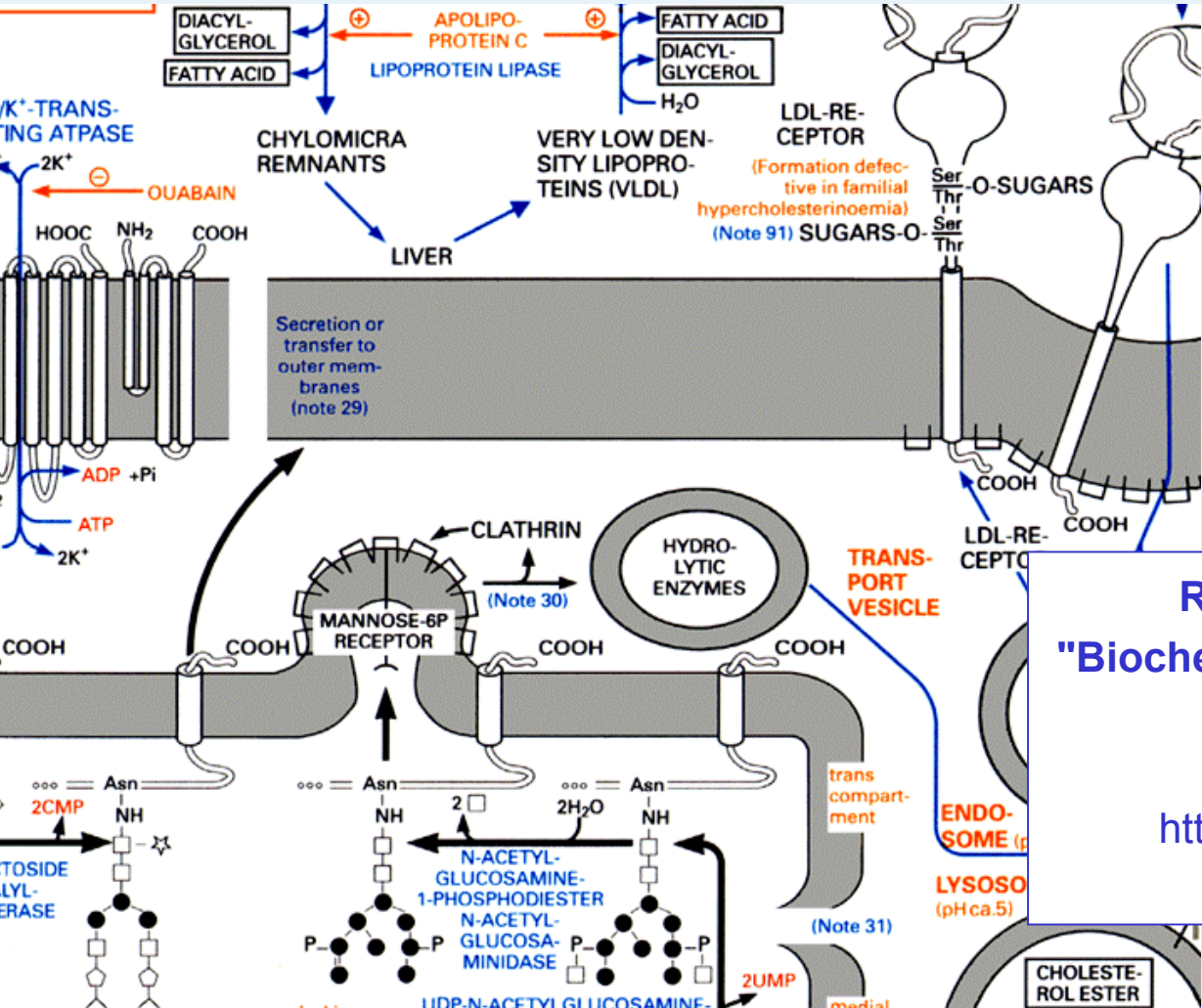
# Proteomics



**2D Gel von *E.coli* K-12 W3110**

<http://www.expasy.org/ch2dgifs/ECOLI/ECOLI.gif>

# Metabolomics

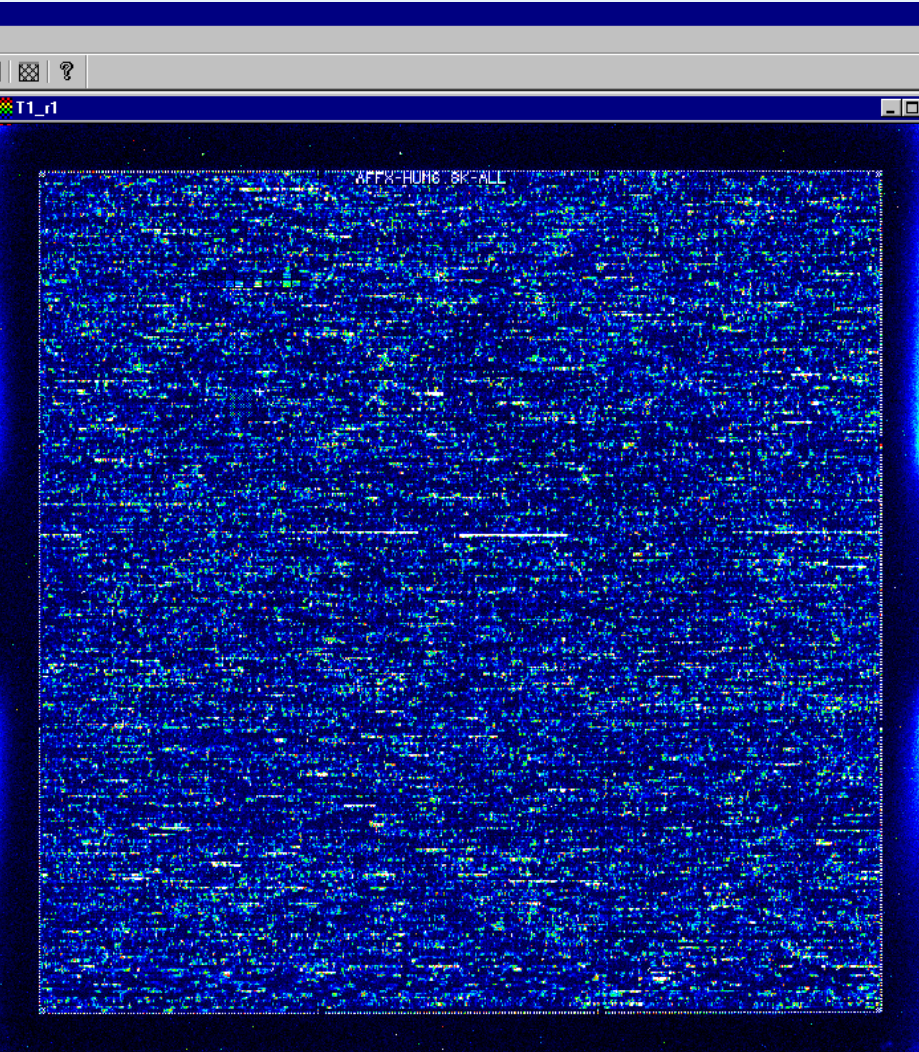


Roche Applied Science  
 "Biochemical Pathways" wall chart

[http://www.expasy.org/cgi-bin/show\\_image?S3](http://www.expasy.org/cgi-bin/show_image?S3)



# Transcriptomics



**Microarray**

**(GeneChip expression arrays)**

<http://www.fgcz.ch/index.php?toc=30>

# Biologische Daten

Publikationen : Information in Freitext; Stichwörter: Autor, Journal, ...

Datenbanken : wohldefiniertes Datenmodell (z.B. Sequenzdatenbank)

DB-System: ASCII Text oder relationale Datenbank

"discovery driven research": Erzeugung ...

- großer Mengen
- neuer Datentypen
- permanente Änderung der Datenmodelle



## Die Grenzen einer Publikation

*Nature* **409**, 942 - 943 (2001)

### **The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X**

D. R. BENTLEY\*, P. DELOUKAS\*, A. DUNHAM\*, L. FRENCH\*, S. G. GREGORY\*, S. J. HUMPHRAY\*, A. J. MUNGALL\*, M. T. ROSS\*, N. P. CARTER\*, I. DUNHAM\*, C. E. SCOTT\*, K. J. ASHCROFT\*, A. L. ATKINSON\*, K. AUBIN\*, D. M. BEARE\*, G. BETHEL\*, N. BRADY\*, J. C. BROOK\*, D. C. BURFORD\*, W. D. BURRILL\*, C. BURROWS\*, A. P. BUTLER\*, C. CARDER\*, J. J. CATANESE†, C. M. CLEE\*, S. M. CLEGG\*, V. COBLEY\*, A. J. COFFEY\*, C. G. COLE\*, J. E. COLLINS\*, J. S. CONQUER\*, R. A. COOPER\*, K. M. CULLEY\*, E. DAWSON\*, F. L. DEARDEN\*, R. M. DURBIN\*, P. J. DE JONG†, P. D. DHAMI\*, M. E. EARTHROWL\*, C. A. EDWARDS\*, R. S. EVANS\*, C. J. GILLSON\*, J. GHORI\*, L. GREEN\*, R. GWILLIAM\*, K. S. HALLS\*, S. HAMMOND\*, G. L. HARPER\*, R. W. HEATHCOTT\*, J. L. HOLDEN\*, E. HOLLOWAY\*, B. L. HOPKINS\*, P. J. HOWARD\*, G. R. HOWELL\*, E. J. HUCKLE\*, J. HUGHES\*, P. J. HUNT\*, S. E. HUNT\*, M. IZMAJLOWICZ\*, C. A. JONES\*, S. S. JOSEPH\*, G. LAIRD\*, C. F. LANGFORD\*, M. H. LEHVASLAIHO\*, M. A. LEVERSHA\*, O. T. MCCANN\*, L. M. MCDONALD\*, J. MCDOWALL\*, G. L. MASLEN\*, D. MISTRY\*, N. K. MOSCHONAS‡, V. NEOCLEOUS§, D. M. PEARSON\*, K. J. PHILLIPS\*, K. M. PORTER\*, S. R. PRATHALINGAM\*, Y. H. RAMSEY\*, S. A. RANBY\*, C. M. RICE\*, J. ROGERS\*, L. J. ROGERS\*, T. SARAFIDOU‡, D. J. SCOTT\*, G. J. SHARP\*, C. J. SHAW-SMITH\*, L. J. SMINK\*, C. SODERLUND\*, E. C. SOTHERAN\*, H. E. STEINGRUBER\*, J. E. SULSTON\*, A. TAYLOR\*, R. G. TAYLOR\*, A. A. THORPE\*, E. TINSLEY\*, G. L. WARRY\*, A. WHITTAKER\*, P. WHITTAKER\*, S. H. WILLIAMS\*, T. E. WILMER\*, R. WOOSTER\* & C. L. WRIGHT\*

\* The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

‡ Department of Biology, University of Crete and Institute of Molecular Biology and Biotechnology, PO Box 2208, 71409 Heraklion, Crete, Greece

§ Neurogenetic Laboratory, The Cyprus Institute of Neurology and Genetics, 6, International Airport Avenue, PO Box 23462, 1683 Nicosia, Cyprus

† Children's Hospital-BACPAC Resources, 747 52nd Street, Oakland, California 94609, USA

# Publikation der Ergebnisse in Datenbanken

**Cross-references**

EMBL	M38352; AAA33878.1; -. [ <a href="#">EMBL</a> / <a href="#">GenBank</a> / <a href="#">DDBJ</a> ] [ <a href="#">CoDingSequence</a> ] D13206; BAA02493.1; -. [ <a href="#">EMBL</a> / <a href="#">GenBank</a> / <a href="#">DDBJ</a> ] [ <a href="#">CoDingSequence</a> ] S39525; AAC60540.2; -. [ <a href="#">EMBL</a> / <a href="#">GenBank</a> / <a href="#">DDBJ</a> ] [ <a href="#">CoDingSequence</a> ] D12680; BAA02181.1; -. [ <a href="#">EMBL</a> / <a href="#">GenBank</a> / <a href="#">DDBJ</a> ] [ <a href="#">CoDingSequence</a> ]
PIR	JQ1390; JQ1390.
PDB	1LGY; 23-DEC-96. [ <a href="#">ExpASy</a> / <a href="#">RCSB</a> ] 1TIB; 26-JAN-95. [ <a href="#">ExpASy</a> / <a href="#">RCSB</a> ] 1TIC; 26-JAN-95. [ <a href="#">ExpASy</a> / <a href="#">RCSB</a> ]
InterPro	<a href="#">IPR000734</a> ; Lipase. <a href="#">IPR002921</a> ; Lipase_3. <a href="#">Graphical view of domain structure.</a>
Pfam	<a href="#">PF01764</a> ; Lipase_3; 1.
PROSITE	<a href="#">PS00120</a> ; LIPASE_SER; 1.
ProDom	[ <a href="#">Domain structure</a> / <a href="#">List of seq. sharing at least 1 domain</a> ].
BLOCKS	<a href="#">P21811</a> .
DOMO	<a href="#">P21811</a> .
ProtoMap	<a href="#">P21811</a> .
PRESAGE	<a href="#">P21811</a> .
DIP	<a href="#">P21811</a> .
SWISS-2DPAGE	<a href="#">GET REGION ON 2D PAGE.</a>

**Keywords**

[Hydrolase](#); [Lipid degradation](#); [Signal](#); [3D-structure](#).

**Features**

Key	From	To	Length	Description
-----	------	----	--------	-------------

DNA-Datenbank

Struktur-Datenbank


Motif-Datenbanken

NiceProt View of SWISS-PROT: P21811 - Netscape

File Edit View Go Communicator Help

### Features

Key	From	To	Length	Description
SIGNAL	<a href="#">1</a>	<a href="#">26</a>	26	POTENTIAL.
CHAIN	<a href="#">27</a>	<a href="#">392</a>	366	LIPASE.
ACT_SITE	<a href="#">268</a>	<a href="#">268</a>		CHARGE RELAY SYSTEM (BY SIMILARITY).
ACT_SITE	<a href="#">327</a>	<a href="#">327</a>		CHARGE RELAY SYSTEM (BY SIMILARITY).
ACT_SITE	<a href="#">380</a>	<a href="#">380</a>		CHARGE RELAY SYSTEM (BY SIMILARITY).
DISULFID	<a href="#">152</a>	<a href="#">391</a>		
DISULFID	<a href="#">163</a>	<a href="#">166</a>		
DISULFID	<a href="#">358</a>	<a href="#">367</a>		
CONFLICT	<a href="#">348</a>	<a href="#">348</a>		I -> M (IN REF. 3).

 [Feature table viewer](#)

### Sequence information

Length: **392 AA** [This is the length of the unprocessed precursor]    Molecular weight: **42138 Da** [This is the MW of the unprocessed precursor]    CRC64: **D08F651EE77AA5A3** [This is a checksum on the sequence]

10	20	30	40	50	60
MVSFISISQG	VSLCLLVSSM	MLGSSAVPVS	GKSGSSNTAV	SASDNAALPP	LISSRCAPPS
70	80	90	100	110	120
NKGSKSDLQA	EPYMQKNT	WYESHGGNLT	SIGKRDDNLV	GGMTLDLPSD	APPISLSST
130	140	150	160	170	180
NSASDGGKVV	AATTAQIQEF	TKYAGIAATA	YCRSVVPGNK	WDCVQCQKWW	PDGKIITFT

Annotation

Sequenz

# Anwendungen

- Diagnostik
- Pharmacogenomics (Korrelation Genotyp-Phänotyp)
- Protein Engineering
- Metabolic Engineering

## Welche Daten werden erzeugt und ausgetauscht?

- Experimentelle Rohdaten oder biologisch interpretierte Daten  
Informationen über Versuchsbedingungen (LIMS)  
Ergebnisse (ASCII, Bildmaterial)
- Textbasierte Interpretation
- Programme zur Auswertung
  
- Austausch: World Wide Web
- Problem: Standardisierung (Format; Datenmodell)

# Datenintegration

Datenbanken: primäre, sekundäre Daten



Links zu Einträgen in anderen Datenbanken (Bsp: SwissProt)



Integration von Datenbanken



Data Warehouse

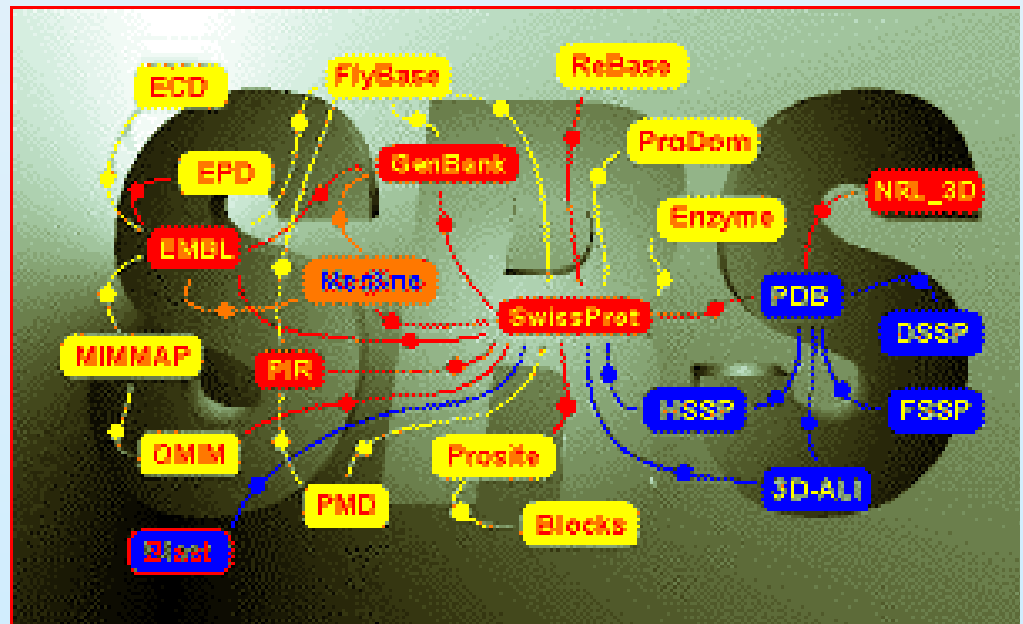
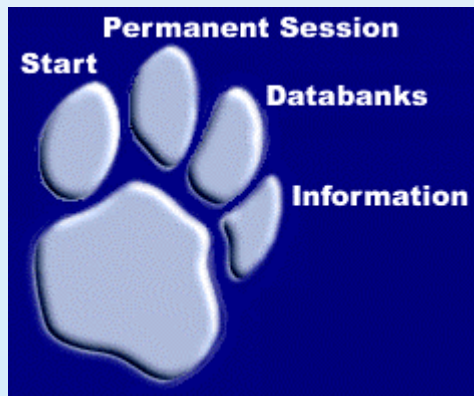
Federated Database System

BioXML



# Datenintegration : das Sequence Retrieval System

<http://srs.ebi.ac.uk/> → Integrierter Zugang zu 140 Datenbanken



# Verwendung der Daten

## Forschung

- Verknüpfung mit eigenen Daten
  - Analyse von (fremden) Rohdaten unter unterschiedlichen Aspekten
- Theoriebildung

## Lehre (Vorlesung, Praktika)

- zusätzlich zu klassischem Lehrmaterial (Lehrbücher, Skripte): interaktives und über das Internet zugängliches Lehrmaterial
- Möglichkeit, im Rahmen von Praktika auf Originaldaten(-banken) zuzugreifen (nicht nur reduzierte Versuchsaufbauten)

# Wie wird publiziert?

## Publikation von Forschungsergebnissen

- Publikation in "peer-reviewed journals" (impact factor!); zusätzlich:
- supplemental material (Daten, Programme) auf server des Verlags
- zusätzliche Daten auf Anfrage bei den Autoren
- Verweis auf eigene homepage (Problem: keine Langzeitsicherung)
- Einträge in externen, "offiziellen" Datenbanken (z.B. Sequenz, Struktur)
- Betrieb einer eigenen, langfristigen Datenbank (commitment)

Diplom- und Doktorarbeiten: ausschließlich Druckversion oder E-Print

# Wie werden Daten gesucht?

## Literatur-Recherche (Medline, CA)

nach Stichwort oder Autor

wer hat eine gegebene Arbeit zitiert?

Volltext-Zugang: online journal

lokale Bibliothek

SUBITO

Suche in online-Angeboten:

Verweise in Publikationen oder home pages

Portale von Fachgruppen oder Arbeitsgruppen

allgemeine Suchmaschinen (Google)

## Trends

- Literatur muß online zugänglich sein
- Zunahme der Relevanz von Rohdaten
- home pages der Forschergruppen zusätzlich zu "offiziellen" Datenbanken

Problem: Datenbereitstellung über langen Zeitraum (Volatilität der Daten)

Zugänglichkeit der Daten / Datensicherheit

